Integrating Data for Analysis, Anonymization, and SHaring

NLP Virtual Machine Manual

Document Version: 1.0.1

University of California, San Diego
Division of Biomedical Informatics

Table of Contents

About this Guide 3
1. Installing the VirtualBox 4
2. Installing the NLP Virtual Machine 4
3. Executing the NLP Virtual Machine 6
4. Demo 8
   eHOST 8
   MIST 9
   MetaMap 10
   cTAKES 11
   ARC 13
5. References 14

About this Guide

This virtual machine (VM) provided by iDASH has a suite of open-source NLP tools installed:

- eHOST - an annotation interface for manually annotating text
- MetaMap - a tool for indexing text to UMLS concepts
- cTAKES – a UIMA-based information extraction system
- MIST - a suite of tools for identifying and redacting personally identifiable information (PII) in free-text medical records
- Automated Retrieval Console (ARC) - a tool for creating a classifier for text based on NLP-derived features

This guide will help you to learn about where to download and how to install this VM. It also provides some demos on how to use the installed NLP tools. To use the VM, you will need your Unified Medical Language System® (UMLS) username and password. If you do not have one already, go to the UTS website to request one.

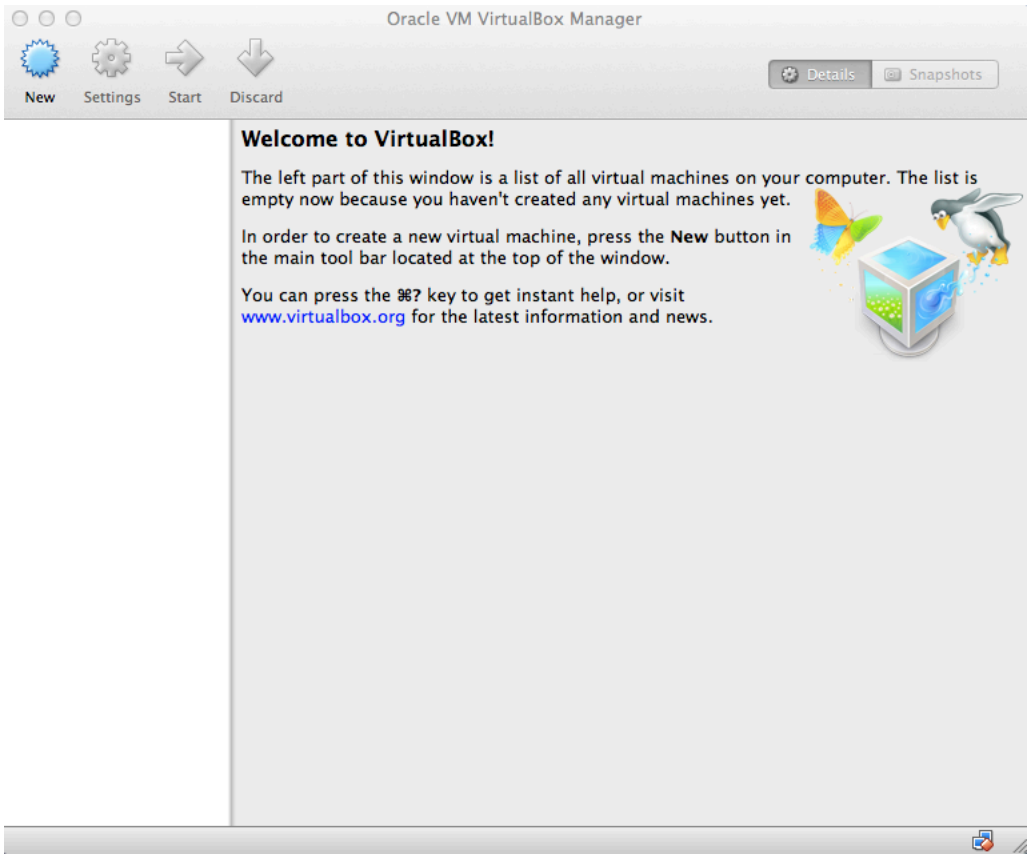
The NLP VM and this manual were created and compiled by Ruiling Liu, an intern at Division of Biomedical Informatics, UC San Diego under the 2014 iDASH summer internship program.

1. Installing the VirtualBox

1) Download VirtualBox from https://www.virtualbox.org/wiki/Downloads
2) Install the VirtualBox

2. Installing the NLP Virtual Machine

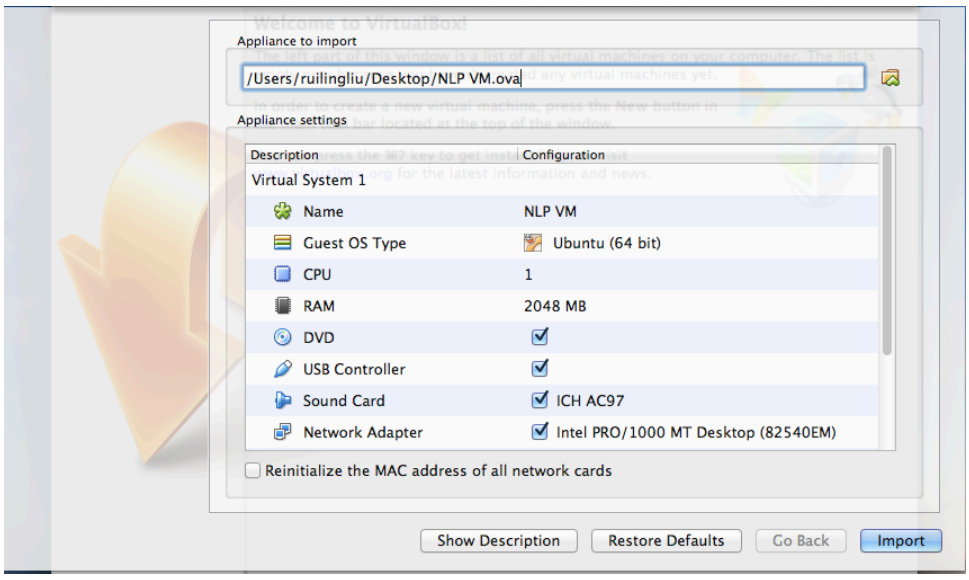
1) Download NLP virtual machine from http://idash.ucsd.edu/nlp/umls-vm
2) Open VirtualBox


Oracle VM VirtualBox Manager
New
Settings
Start
Discard
Details
Snapshots
Welcome to VirtualBox!
The left part of this window is a list of all virtual machines on your computer. The list is empty now because you haven't created any virtual machines yet.
In order to create a new virtual machine, press the New button in the main tool bar located at the top of the window.
You can press the ⌘? key to get instant help, or visit www.virtualbox.org for the latest information and news.

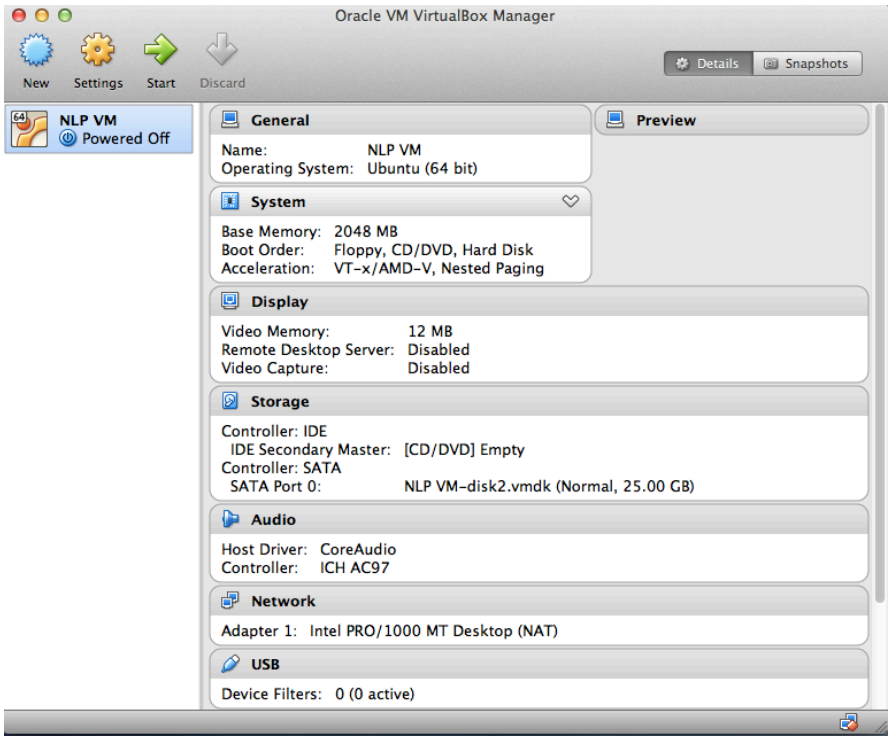
3) Import the NLP Virtual Machine

The downloaded NLP VM is a single archive file with the .ova extension. To import the "NLP VM.ova", select "File" -> "Import appliance" from manager window. In the file dialog that comes up, navigate to the file "NLP VM.ova".

A dialog similar to the following will appear:

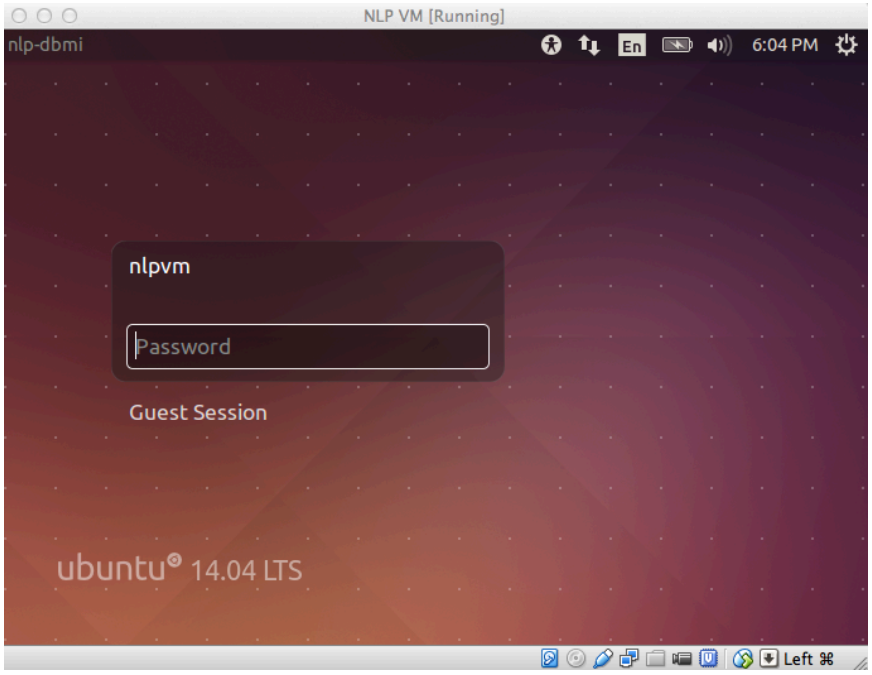


This presents the virtual machine described in "NLP VM.ova" and allows you to change the virtual machine settings by double-clicking on the description items. Once you click on "Import", VirtualBox will copy the disk image and create a local virtual machine with the settings described in the dialog. The imported virtual machine will show up in the Manager's list of virtual machines.



3. Executing the NLP Virtual Machine

1) Click on “Start” to start the NLP virtual machine

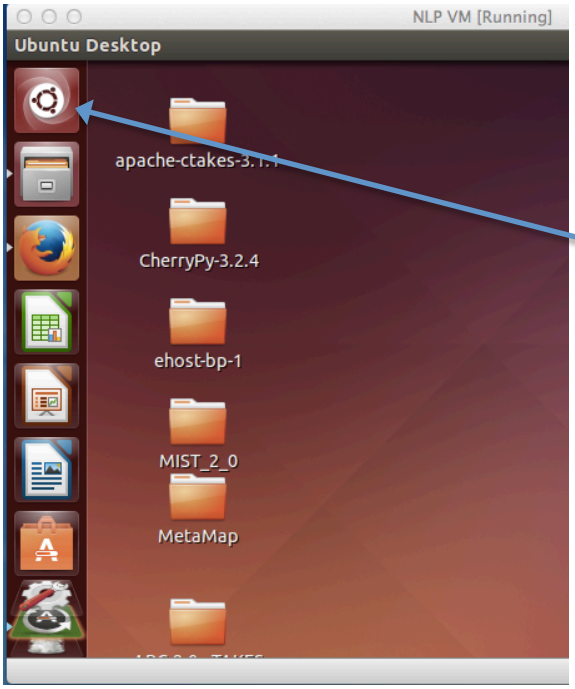


2) Login the NLP VM

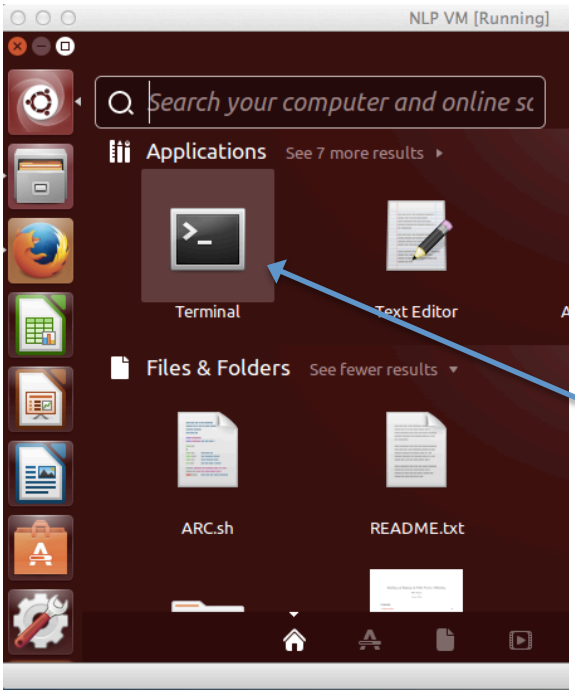
The username for this NLP virtual machine is **nlpvm**, and the password is **dbmi123**.

3) Open the terminal

Click on the icon in the top left corner of the desktop and then click on the terminal icon to open the terminal



Click on the icon in the top left corner of the desktop



Click on the terminal icon to open the terminal

4. Demo

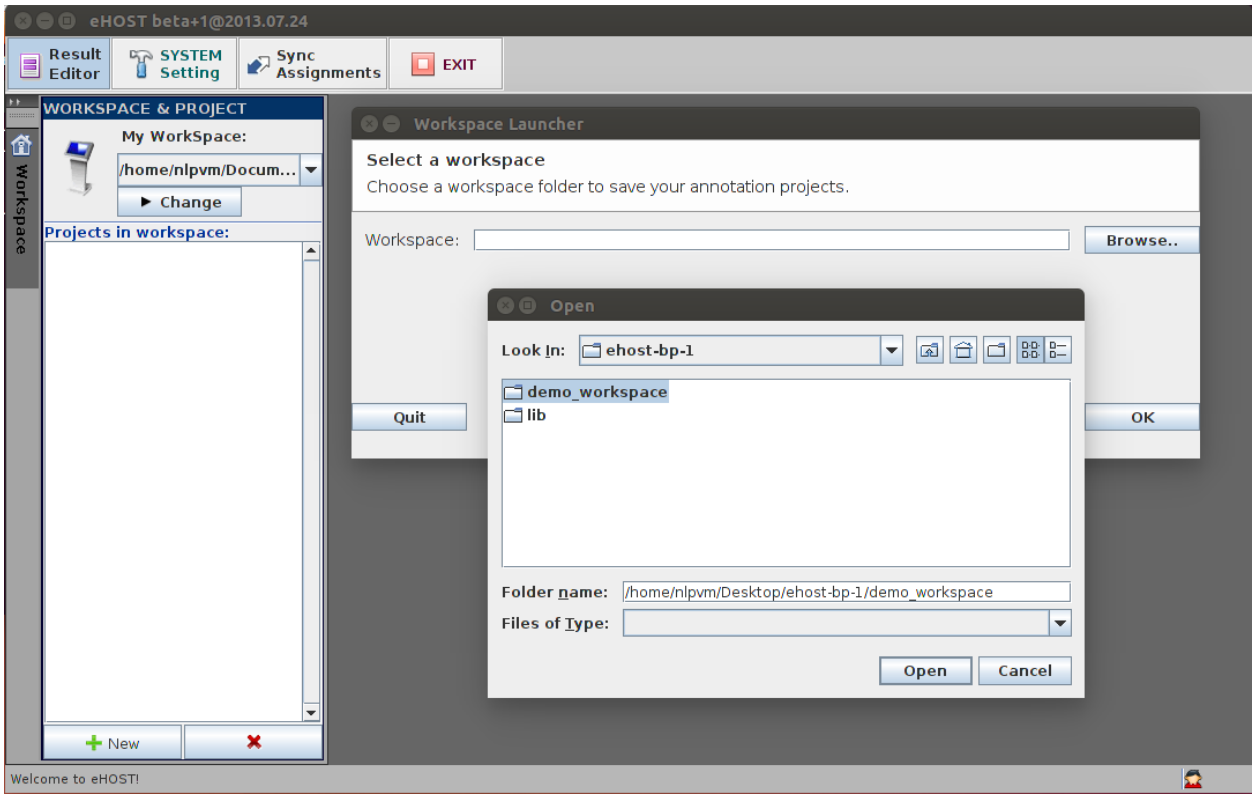
eHOST

1. Open terminal
2. Enter the command below in the terminal to open eHOST

   $ cd /home/nlpvm/Desktop/ehost-bp-1

   $ java –jar eHOST.jar
3. Click “Brower” button to navigate to the demo_workspace folder and click on “Open”

   The location of the demo_workspace is /home/nlpvm/Desktop/ehost-bp-1/demo_workspace
4. Click on “OK”



5. Use the official document of eHOST and the demo data to explore the functions of eHOST: http://ehost.knowtator.com/html/start.html

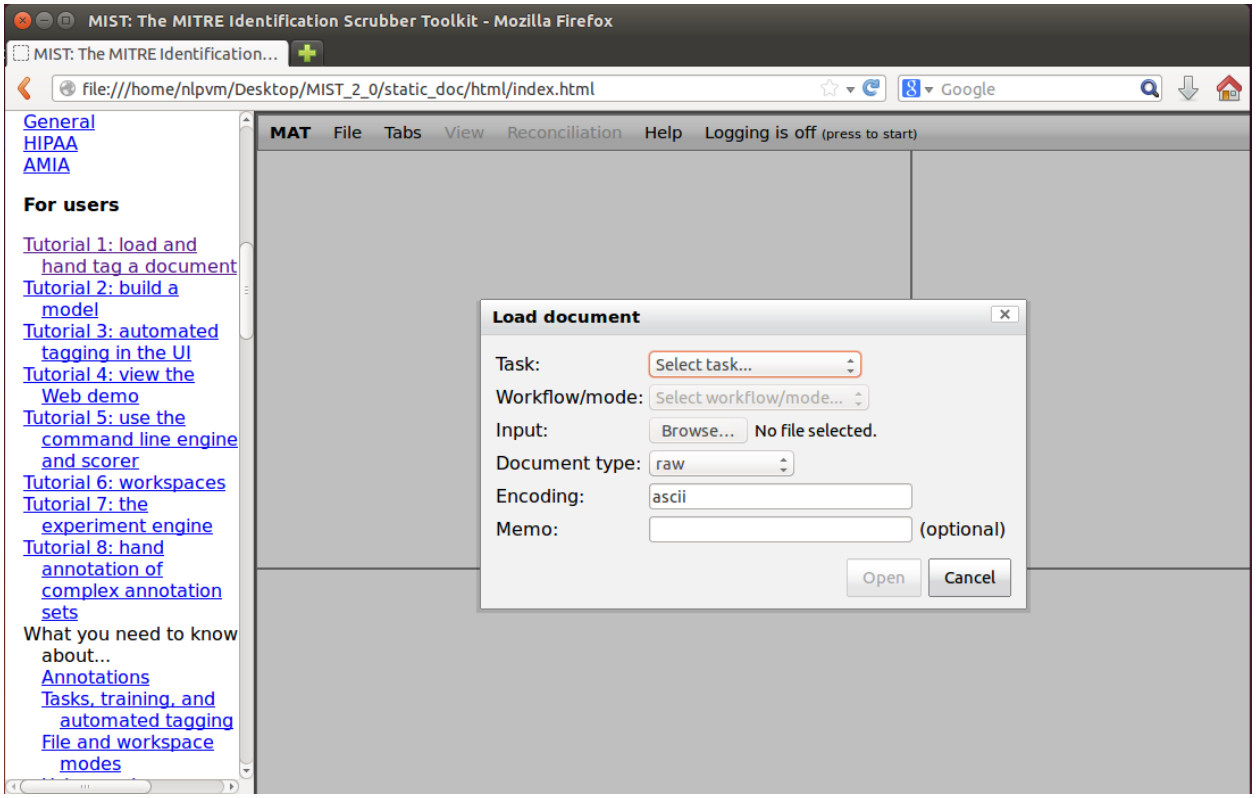
MIST

1. Open the terminal
2. Install the task

   The task has already been installed under the folder in the NLP VM:

   /home/nlpvm/Desktop/MIST_2_0/src/MAT
3. Start the UI by entering the command in the terminal

   $ bin/MATWeb
4. Open your Firefox browser and:
   a. Ensure that popups are not blocked for localhost (you can check this in the Firefox preferences window under the content tab)
   b. In the Firefox preferences window in the General tab, ensure that "Always ask me where to save files" is selected.
5. Navigate to http://localhost:7801/MAT/workbench


MIST: The MITRE Identification Scrubber Toolkit - Mozilla Firefox
MIST: The MITRE Identification...
file:///home/nlpvm/Desktop/MIST_2_0/static_doc/html/index.html
Google
General
HIPAA
AMIA
For users
Tutorial 1: load and hand tag a document
Tutorial 2: build a model
Tutorial 3: automated tagging in the UI
Tutorial 4: view the Web demo
Tutorial 5: use the command line engine and scorer
Tutorial 6: workspaces
Tutorial 7: the experiment engine
Tutorial 8: hand annotation of complex annotation sets
What you need to know about...
Annotations
Tasks, training, and automated tagging
File and workspace modes
MAT File Tabs View Reconciliation Help Logging is off (press to start)
Load document
Task: Select task...
Workflow/mode: Select workflow/mode...
Input: Browse... No file selected.
Document type: raw
Encoding: ascii
Memo: (optional)
Open
Cancel

6. Use the official document of MIST to explore the other functions of MIST

   file:///home/nlpvm/Desktop/MIST_2_0/static_doc/html/index.html

MetaMap

1. Open the terminal

MetaMap requires the starting of two servers. The Part-of-Speech Tagger and Word Sense Disambiguation (WSD) Servers. Both servers will automatically run in the background when started. They can be started and stopped as follows:

2. Start the SKR/Medpost Part-of-Speech Tagger Server by entering the command below in the terminal

$ cd /home/nlpvm/Desktop/MetaMap/public_mm

$ sudo ./bin/skrmedpostctl start

$ [sudo] password for nlpvm is: dbmi123

3. Start the Word Sense Disambiguation (WSD) Server by entering the command:

$ sudo ./bin/wsdserverctl start

4. Using MetaMap

$ echo “lung cancer”|./bin/metamap11v2 -I

5. Stop the SKR/Medpost Part-of-Speech Tagger Server

$ sudo ./bin/skrmedpostctl stop

6. Stop the Word Sense Disambiguation (WSD) Server

$ sudo ./bin/wsdserverctl stop

7. Use the official document of MIST to explore the other functions of MetaMap

http://metamap.nlm.nih.gov

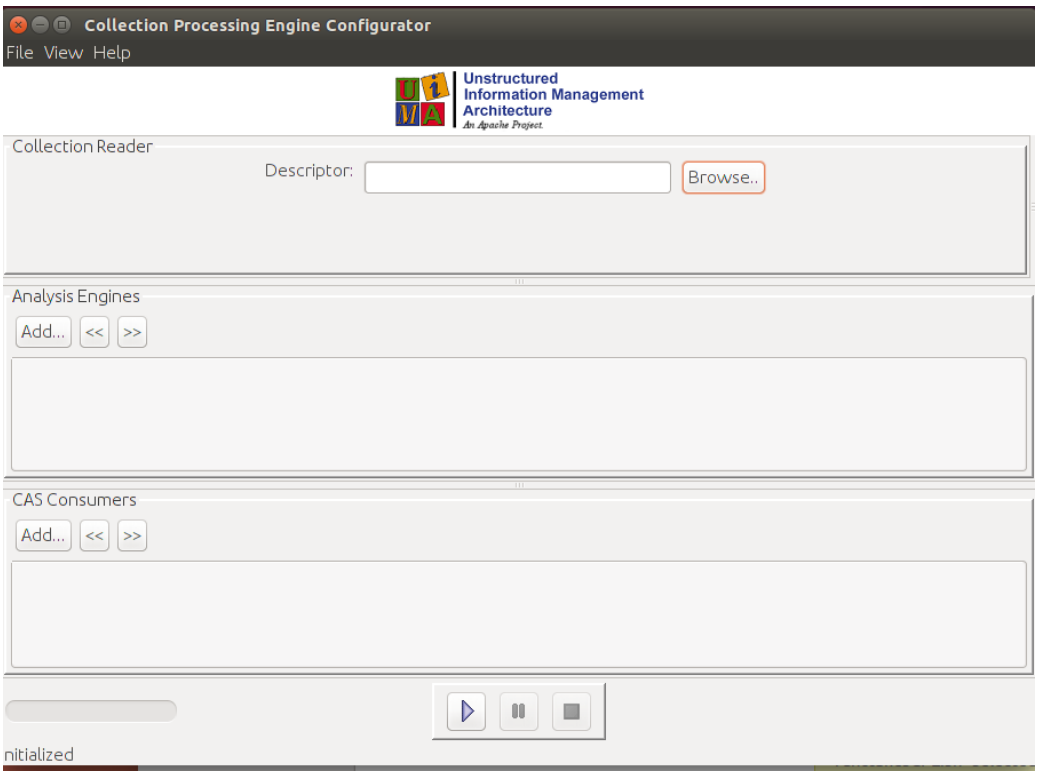
cTAKES

In the initial setup cTAKES will recognize only few sample concepts in text. If you wish to perform named entity recognition or concept identification for anything other than these few words, you will need to 1) obtain the rights to use UMLS resources 2) add those credentials to cTAKES, and 3) use an aggregate that makes use of those UMLS resources. If you don't, cTAKES will work but won't recognize much.

1. Open terminal
2. Enter the commands below in the terminal to open cTAKES

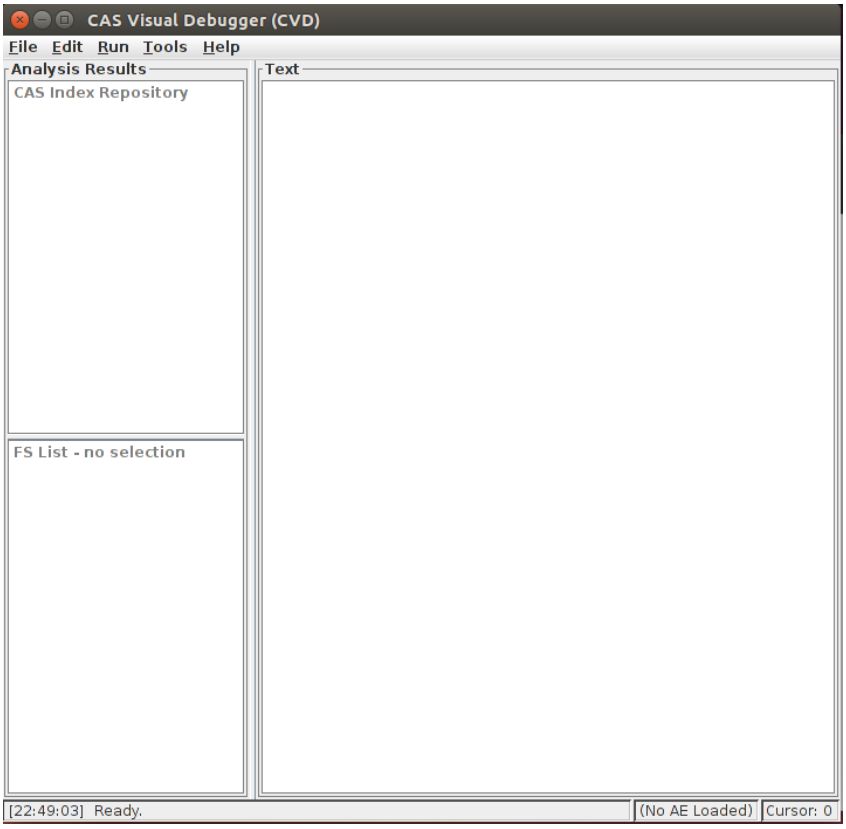
   $ cd /home/nlpvm/Desktop/apache-ctakes-3.1.1/bin

   $ sh runctakesCPE.sh



3. Enter the command below to open the CAS Visual Debugger (CVD)

   $ sh runctakesCVD.sh



4. (Optional) Add UMLS access rights
    a. If you do not have a UMLS username and password, you may request one at UMLS Terminology Services.
    b. Edit the following files. Find the line in each script that runs java and add the ctakes.umlsuser and ctakes.umlspw parameters to the java command with your credentials. Make sure you substitute your actual ID and password if you cut and paste the example.

Files:
<cTAKES_HOME>/bin/runctakesCVD.sh
<cTAKES_HOME>/bin/runctakesCPE.sh
Java Command:

java -Dctakes.umlsuser=<YOUR_UMLS_ID_HERE> -Dctakes.umlspw=<YOUR_UMLS_PASSSWORD_HERE> -cp ...

5. Use the official document of cTAKES to explore the functions of cTAKES https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.1+User+Install+Guide%3bjsessionid=0F37630963F4482D9A9AD64578F9E9CD#cTAKES3.1UserInstallGuide-CollectionProcessingEngine(CPE)

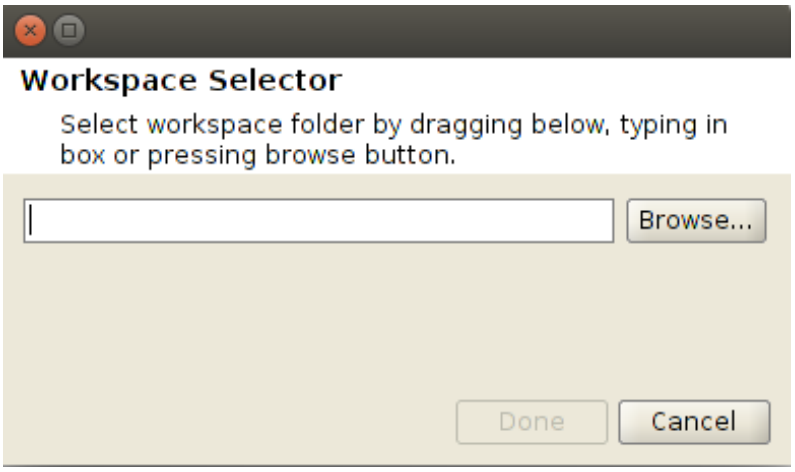
ARC

1. Open the terminal
2. Start the ARC program by entering the commands below:

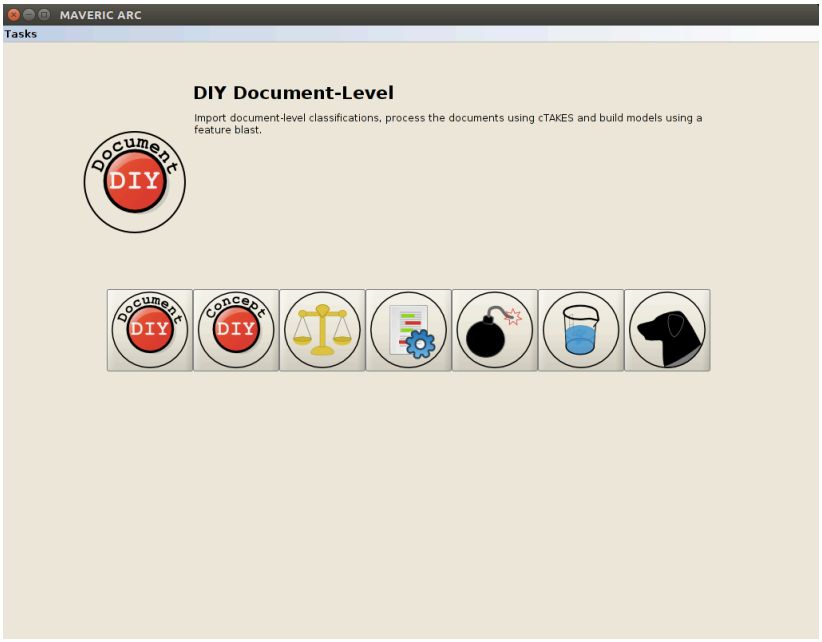
$ cd /home/nlpvm/Desktop/ARC-2.0-cTAKES

$ sh ARC.sh

3. Select workspace



4. Click on “Done”



5. References

1) Oracle VM VirtualBox® User Manual. https://www.virtualbox.org/manual/ch01.html#intro-installing
2) eHOST User Guide. https://code.google.com/p/ehost/
3) cTAKES 3.1 User Install Guide. https://cwiki.apache.org/confluence/display/CTAKES/cTAKES+3.1+User+Install+Guide
4) MetaMap. http://metamap.nlm.nih.gov
5) ARC 2.0 Installer with LVG and UMLS pre-bundled. https://code.google.com/p/mavericarc/
6) The MITRE Identification Scrubber Toolkit. http://mist-deid.sourceforge.net